Testing and optimization of a continuous speech recognizer with a middle sized vocabulary[1]

Csaba Teleki, Szabolcs L. Tóth, and Klára Vicsi

This paper describes testing and optimization methods of a speaker independent, continuous and automatic speech recognizer, developed at the Laboratory of Speech Acoustics. The recognizer is based on acoustic and language models built using statistical methods. It is capable of training and recognizing middle sized vocabulary-based texts, containing 1000-20 000 different words. New methods were developed in acoustical preprocessing [1], in statistical model developing and language modeling. The recognizer uses HMM acoustic models for phonemes [2] and bigram language model, using non-linear smoothing [3]. During the tests we varied the acoustical models and the language model of the recognizer.

The acoustic models were trained, using the Hungarian Reference Database (MRBA) [4].The optimization of these models were done earlier. In the end, the recognizer uses Quasi-Continuous HMM (QCHMM) models with 4-5 states based on recordings made with 16 kHz sample rate, 17 derivates in Bark frequency domain, 17 time derivates, 17 second time derivates, and energy as input vector. For details on optimization of acoustic level see [5].

The language model were developed using a text corpus collected from 2700 medical reports from the of Medical Clinic nr 2 of SOTE in Budapest and 6365 medical reports from the Medical School of Szeged.

For testing purposes we used a couple of different type of recordings. One of them was made at Medical Clinic nr 2 of SOTE in Budapest. The test material contains spoken medical reports of 5 different doctors each physician dictating 4 different medical reports. Therefore, 20 different recordings, from 5 different speakers were used for testing. The recordings were made in an examination room with bad acoustic condition. The other type of testing material was recorded in our laboratory by 2 speakers, containing the same text, but in better acoustic condition.

The recognizer was evaluated from multiple points of view: word error rate (WER), real time factor (RT), memory capacity used during recognition (MEM).

This paper presents results that show that the goodness of the recognizer is influenced by:

- the type training material,
- the acoustic conditions of the testing material,
- the pronunciation quality of the speaker,
- the size of the searching space,
- the weighting of the role of the bigram language model.

We trained three different types of acoustic models: one based only on male speech, another one based on only on female speech and one based on mixed (male and female) speech from the MRBA database. So far, for the recordings made at the Medical Clinic, the word error rate is 30% for lexemes, using the language model built based on the reports made at the Medical Clinic from Budapest. Using the recordings made at our laboratory, in better acoustical conditions, the results got significantly better (WER = 21%). Mean value for the real time factor were 1 and for the memory capacity used were 300MB. The pronunciation quality of the speaker influences the WER. If the pronunciation quality is higher, de recognizer gives better results. If the size of the searching space is bigger, the results will be better, but then increases the processing time and the used memory.

[1]The work has been supported by the Hungarian Scientific Research Foundations (OTKA T 046487 ELE) and by The National Office of Research and Technology (IKTA 00056).

The above mentioned preliminary results showed that optimization of the acoustic and language model of the recognizer is required. Beyond tuning the mentioned parameters to the appropriate values, optimization at different levels of the training material is necessary. At the level of acoustic models we will show results regarding the effectiveness of using supervised speaker adaptation. At the level of the language model we tried to optimize the vocabulary, using perplexity and active vocabulary adaptation. Using these optimization methods, the word error rate decreased under 10 %.

References

[1] M. Szarvas and S. Furui. Evaluation of the Stoachic Morphosyntactic Language Model on a One Million Word Hungarian Dictation Task, In: *EUROSPEECH 2003 - Genova*, p. 2297-2300

[2] S. Chen, D. Beeferman, and R. Rosenfeld. Evaluation Metrics For Language Models, In: *DARPA98*, National Institue of Standards and Technology (NIST)

[3] P. Clarkson, and T. Robinson. Towards improved language model evaluation measures

[4] K. Vicsi, A. Kocsor, Cs. Teleki, and L. Tóth. Hungarian Speech Database for computer-using environments in offices, *Second Conference on Hungarian Computational Linguistics (MSZNY 2004)*, Szeged, Hungary, 2004, p. 315-319.

[5] Sz. Velkei and K. Vicsi. Speech recognizer model-building experiments at the level of acoustic and phonetics, on behalf of developing a speech recognizer for medical reporting, *Second Conference on Hungarian Computational Linguistics (MSZNY 2004)*, Szeged, Hungary, 2004, p. 307-315.

[6] C. Becchetti, L.P. Ricotti. Speech Recognition,*Theory and C++ implementation*, Fondazione Ugo Bordoni, Rome, 1999. ISBN 0-471-97730-6

[7] HUMOR, Morphological analyzer for developers,
`http://www.morphologic.hu/h_humor.htm`